數位語音處理概論期末專題

所選題目：論文閱讀

所選論文：

***Unsupervised Discovery of Linguistic Structure Including Two-Level Acoustic Patterns Using Three Cascaded Stages of Iterative Optimization***

Cheng-Tao Chung, Chun-an Chan, and Lin-shan Lee

教授：李琳山

學生：鐘民憲(B06901017)

1. 論文內容

摘要

雖然現今大量的語音資料唾手可得，但是要人為去一一標記這些資料實在太過勞力費時，因此非監督式辨識聲學特徵的技術顯得越來越重要，這篇論文提出了一個對於目標語言，能夠直接從原始的語音資料非監督式地找出語言學結構的方法。語言學結構包含兩個層級的聲學特徵：subword-like和word-like(某種程度上可以想成是注音跟字)，在辭典中一個word-like pattern由數個連續的subword-like pattern組成，subword-like pattern用HMM來表達，而N-gram語言模型則是由這些word-like pattern組成。所有的特徵、模型和參數都可以自動從未標記的語料庫中學習，學習的歷程分為初始化跟後續三個階段，這三個階段分別迭代式地優化聲學模型、語言模型和辭典，每一次模型訓練和解碼會產生新的標記，辭典和HMM也會因此跟著更新，藉由這個方法，模型的參數和解碼出來的標記每次都會被分別優化，使得語言學結構的知識被逐層學習。這篇論文將所提出的方法在中文新聞的語料庫上進行初步測試，並將其拿來與監督式訓練的模型進行比較，結果顯示這個系統不僅有不錯的表現，亦能應用在現存詞彙量龐大的ASR系統上。

1. 介紹

自動語音辨識中HMM的監督式訓練不只依賴大量聲學資料的蒐集，也需要精確的標記，儘管監督式學習的模型表現良好，但是成本實在過高，且在大部分情況下無法取得有標記的資料。這也是為何有許多人投入在研究如何非監督式地從大量聲學資料中找出聲學特徵，且不需要任何人力標記與相關知識。大部分研究僅限於找到一層phoneme-like的聲學特徵，然而實際上語音訊號包含多層的結構，而這些多層結構對於分析和解讀語音訊號是有所幫助的。

因此這篇論文提出非監督式尋找兩層聲學特徵(subword-like and word-like)的方法，不只是HMM，subword-like pattern的數量和word-like pattern的辭典大小也都可以自動從資料中學習，而更多語言學結構像是N-gram語言模型和word-like pattern辭典則可以直接由語料庫的聲學訊號取得，這是藉由將動態的辭典整合進傳統監督式HMM的訓練過程，然後執行三階段模型和標記之間的迭代式優化來達到，如此一來模型、參數和語言學結構可以從語料庫中一層一層的學習並微調。藉由這個方法，我們可以找出目標語言語料庫中的word-like pattern以及subword-like pattern。

1. 理論

目標：給定語料庫中一段聲學特徵向量序列，希望找到一個代表語言學結構和word-like pattern標記的參數集合，代表subword-like pattern的HMM，代表辭典，代表語言模型。一開始先透過找到初始標記如(1)，在第i次優化，我們可以藉由上一輪找到的訓練出如(2)，然後再用去解讀出如(3)。

以上步驟為初始化，接下來有三個階段(I)(II)(III)分別對HMM、語言模型和辭典進行優化如Fig1，在Fig1中，每一階段訓練的次數分別為、和，當和之間的差異變得夠小時，便可以進入到下個階段。是由EM訓練出來的，而和則是直接由產生。然而並非每個階段都會被用到，細節請見Fig2。

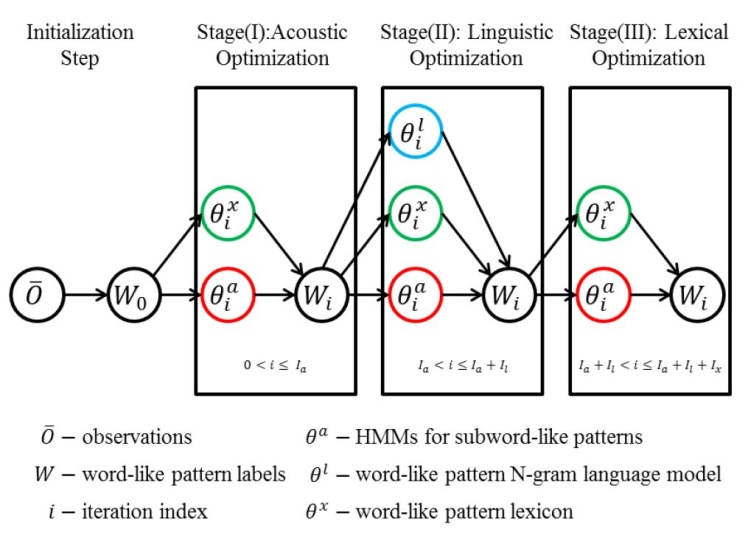


Fig1

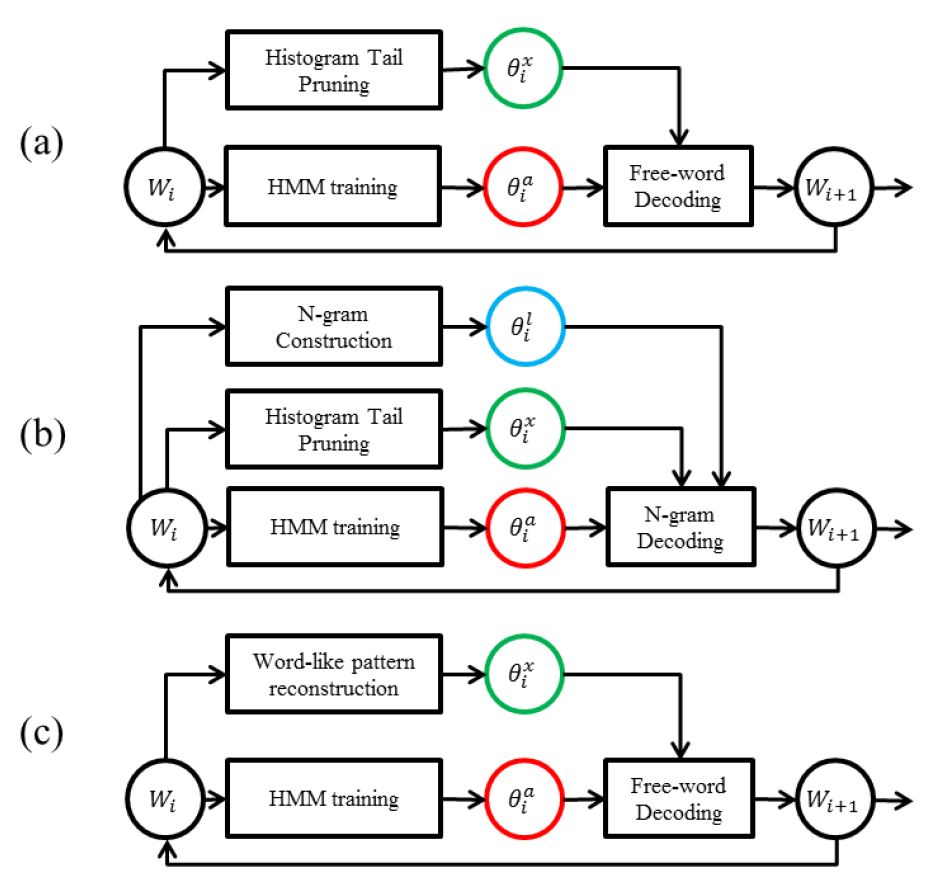


Fig2

Fig1這樣設計流程的初衷為漸進式地調整參數，避免參數卡死在局部最佳解，這樣的情況常發生在很多參數一起優化時。在階段(I)，HMM獨自被訓練因為它們是整個語言學結構的核心區塊，當HMM足夠可信後，在階段(II)使用N-gram參數更好地解讀常常一起出現的word-like pattern，同時持續更新HMM參數，最後在階段(III)將word-like pattern打斷成subword-like pattern，然後再重建出更好的word-like pattern。在階段(I) (II)中辭典word-like pattern的數量可能會因為比較少出現的pattern被其他pattern吸收而減少，但是在階段(III)這個數量會有更明顯的浮動。Subword-like pattern的time alignment在每一輪訓練解碼後都會被更新。

2.1初始化

初始化標記採由上往下的方式，先把每段聲音序列根據MFCC能量大小的不連續性切割成word-like片段，接下來再將每段word-like片段的filtered self-similarity dotplot用watershed transform分割成

subword-like片段，Watershed transform可以在灰階圖像上捕捉物體的數量及邊界，所以在圖中對角線格子與物體的邊界交會處就會被視為subword-like片段的邊界，範例如Fig3。

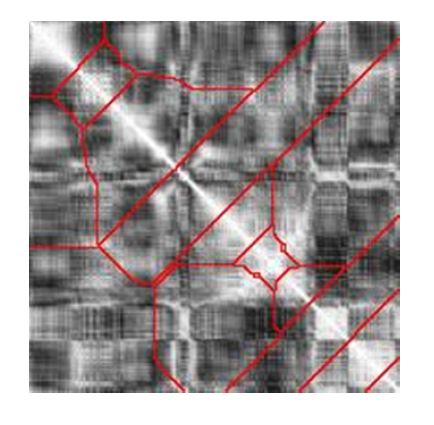


Fig3

接著抽取每段subword-like片段的平均作為代表特徵向量，然後對所有從語料庫中得到的特徵向量執行k-平均演算法，群的總數由群內散布與群間散布的比值決定，最後給每一個群一個subword-like ID。不同subword-like pattern序列決定了word-like pattern，而語料庫中word-like pattern的總數則決定了辭典初始的字彙量，這個語料庫即為。

2.2階段(I)：聲學模型優化

如Fig2(a)，每一次得到的聲學模型集合都是從語料庫經由ML得到的HMMs，辭典則是蒐集所有出現在且出現次數超過閾值的word-like pattern。接著透過和去解讀整個語料庫的，產生新的標記，此時不只中HMM的參數及HMM片段的邊界改變，的字彙量也可能因為某些word-like pattern的出現次數太少而減少。

2.3階段(II)：語言模型優化

如Fig2(b)，這個階段與上一階段相當類似，差別只在於word-like pattern的N-gram語言模型也會從被預測出來，並用在解讀、生成的過程中，N-gram對於生成更好的相當有幫助，尤其當word-like pattern常常一起出現時。

2.3階段(III)：辭典優化

如Fig2(c)，在這個階段，把中的word-like pattern打斷變成subword-like pattern，再根據重建新的word-like pattern。如果某個連續的subword-like pattern序列出現頻率夠高且上下文的變化幅度也夠大，則我們便將其視為一個word-like pattern，這件事情我們可以根據建構PAT-Tree來達成。用這個方法可以發現每一次都會大幅更新，更新過後的便可以用於生成新的標記。當和之間的差異足夠小後，整個訓練就會停止，用最後得到的訓練，全部完成後得到的就是自動找到的語言學結構。

1. 實驗

3.1實驗設置

這個方法初步測試的對象為由2001臺灣廣播新聞長度4小時合計5034個音訊片段組成的中文語料庫。Subword-like pattern的HMMs一共有13個states，每個state都只有一個Gaussian，之所以這樣設定是因為我們假設所關注的subword-like pattern在訊號軌跡的變化量會比聲學上的變化量大，聲學上變化量大的訊號片段應該被分類為不同的patterns。最後，階段(I)(II)(III)都各會對整個語料庫執行30次訓練。

3.2最初的觀察和分析

有趣的是得到的208個subword-like pattern幾乎都可以對應到中文的某個音節，Fig4顯示了208個subword-like pattern與399個人為標記中文音節之間的對應關係，橫軸的中文音節按照聲學上的相似度排序，每個圓圈都代表縱軸上35個或更多的subword-like pattern。雖然鄰近的音節之間有些模糊性，但從此圖仍看出強烈的一對一關係。而362個word-like pattern則對應到大約154個常出現的多音節字和208個單音節字，那些出現頻率不夠高的字無法被分辨，結果便會以一到多個subword-like pattern來表示。

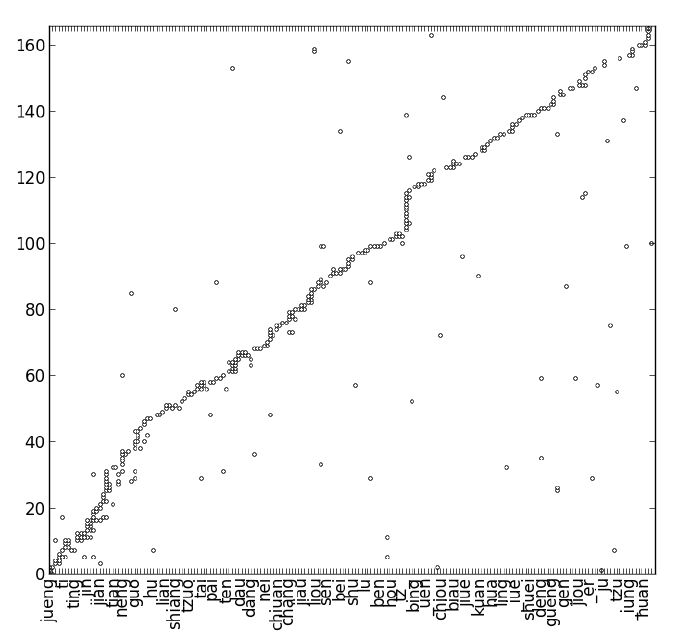


Fig4

Fig5展示了subword-like pattern的數量、word-like pattern的辭典大小、跟的在word-like pattern level上跟utterance level上的一致性，以上三點是如何隨訓練次數改變。總體而言，辭典大小在階段(I)(II)有所下降，而在階段(III)有跳躍和震盪的現象。雖然大部分在階段(I)找到的word-like pattern到階段(II)就會消失，但它們最主要的功能是在訓練subword-like HMM時提供一些上下文的資訊。

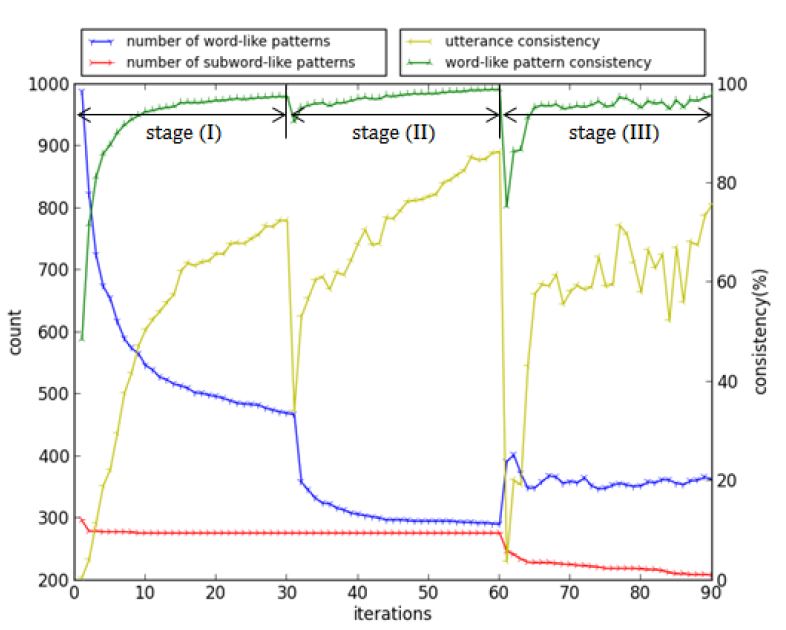


Fig5

3.3實驗驗證初始化和後續三個迭代階段的合理性

進一步的實驗，我們使用與之前提出方法稍微不同的設置，並在語料庫5034個音訊中的942個上進行測試，對於每個設置，我們藉由如Fig4的方法將找到的subword-like pattern與中文音節進行對應去評估音節的正確性。在第一個部分，我們用三種不同的方式產生初始並只執行階段(I)50次。這三種方法分別是：(1)2.1提出的方法，由上而下兩層初始化，從word-like片段開始(2)只在subword上用watershed transform做初始化，不考慮更高層的word-like片段(3)與(2)相似但不做k-平均演算法，相同數量的subword-like pattern ID隨機指定給每個subword-like片段。方法(1)跟(2)最主要的差異在兩層的特徵結構，方法(1)讓我們在初始化的時候就已經在找兩層特徵(subword-like跟word-like)的半途了，方法(2)則比較像是只為了找一層特徵(phoneme-like)而採行的非監督式初始化方法。結果列在Table1左半邊，雖然方法(1)只比方法(2)好了1.03%，但是針對方法(1)產生的pattern進行人工測試顯示出這樣的進步是重要的。

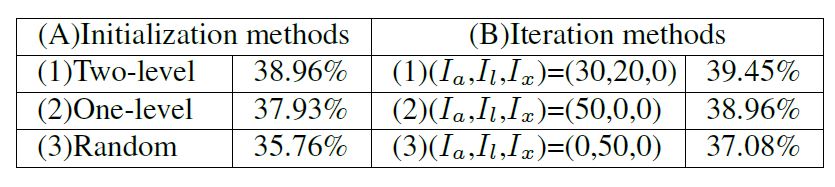


Table1

以上證明了word-like pattern的限制對優化過程是有幫助的，而方法(1)相對於方法(3)隨機指定ID而未經分群的做法正確率高上不少也代表階段(I)的優化訓練是有幫助的。

第二部分，我們用方法(1)產生初始然後再分別採用三種不同的訓練次數：(1) (2) (3)，第一種做了半套，而第三種則是很直覺地同時優化聲學模型和語言模型參數。結果列在Table1右半邊，第一種比第三種好2.37%，而第一種比第二種好代表中間的過渡期是進步的來源，證實了一層一層漸進式地學習可以產出更可信的結果。階段(III)辭典優化的好處在另一篇ICASSP2013口語內容語意恢復的論文可以更容易看出，因為word-like pattern是帶有語意的。

3.4口語詞彙偵測

我們也將這些找出來的patterns應用在口語詞彙偵測上，並與在24.5小時中文廣播新聞人工標示語料庫上訓練出來，trigram語言模型有著72k詞彙，用來做語音辨識的中文音節模型進行比較。我們在相同的情境下測試監督式HMMs與非監督式HMMs的表現，並以監督式HMMs的表現作為上界。我們詢問的52個詞彙包含國家名稱、組織及政治領袖，每次詢問我們都會解讀出它在語料庫中對應的音訊，然後選擇出現頻率最高的HMM序列來代表它，監督式使用的是音節HMMs，而非監督式則是使用subword-like pattern HMMs，接著這個選出的HMM序列會拿去與語料庫中的所有音訊的HMM序列比較，然後以關聯性來評分。第一步先計算每一對離線HMMs之間的距離，在此兩個HMMs的距離是以state序列的DTW-distance定義的，一個HMM中的一個state可以對應到另一個HMM的多個states，反之亦然，DTW的距離矩陣是兩個高斯矩陣的KL-散度。第二步計算選出的HMM序列與線上語料庫HMM序列的距離，在此兩個HMMs序列的距離定義為兩個序列中對應的HMM模型之間的距離總和。因為大部分計算都是離線完成，所以這個方法執行的速度跟文本資訊復原的速度相同。

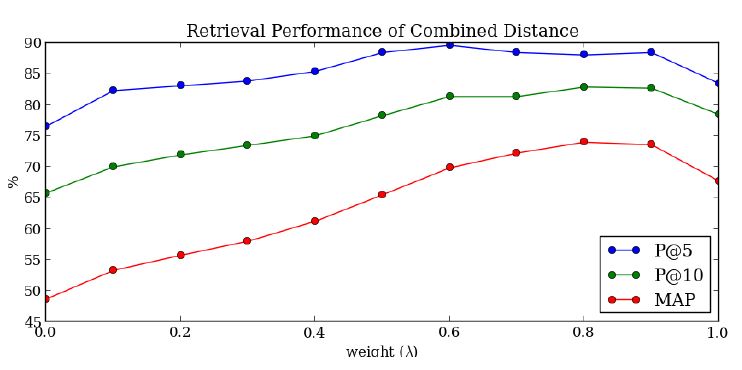


Fig6

監督式總和距離為，非監督式距離總合為，接著利用兩者結合的距離來執行口語詞彙偵測，結果如Fig6，當代表只靠非監督式模型時，表現仍然不錯，更重要的是，兩者結合的距離在三種測試下皆有更好的表現，比起只用監督式或非監督式距離而言，這代表我們提出的方法能夠直接從那些在監督式語音辨識學習中被遺漏的資料中擷取資訊，換句話說，這個方法不只自己的表現好，還能夠與標準的監督式ASR系統相輔相成

1. 結論

本研究提出了一個方法能夠從語料庫中非監督式地找出包含兩層聲學特徵的語言學結構，與之前研究不同之處在於找到的是兩層聲學特徵，以及模型的參數是經過數個階段的迭代式優化，一層一層漸進式學習產生的。雖然一些早期的方法也有將階層式的知識納入考量，但本研究使用的是13個state的單一Gaussian HMM，相較於state數量少、多Gaussian的傳統HMM，能夠接受的聲學變化量相對較少。初步subword-like pattern序列在口語詞彙偵測上的實驗顯示本研究提出的系統可以與ASR系統相輔相成，而另一篇ICASSP2013的論文做了更完整的口語詞彙偵測實驗，顯示此模型表現得比segmental DTW方法更好。此外，第二層word-like pattern的主要目標在於捕捉聲學訊號中的語意特徵，這點在另一篇ICASSP2013口語內容語意恢復的論文中被驗證。

1. 心得

一開始想選擇這篇論文是因為在做專題研究時，那時候訓練HMM模型，都需要事先準備好質跟量都相當不錯的訓練資料，做出來的辨識系統正確率才會高，那時就很好奇，雖然音訊檔案可以相對容易地從新聞報導之類的來源取得，但是那些對應到這些音訊的「正確答案」，也就是文字的部分，要如何生成或取得，除非當初新聞就有字幕，不然如果是用人力一個個檔案聽，然後轉換成文字，真不知道幾萬個檔案要花上多少時間。於是乎就有人開始思考是否有可能在沒有「正確答案」的情況下，照樣訓練出不錯的模型，也就是所謂的「非監督式」學習，如果「非監督式」的成效能夠跟「監督式」分庭抗禮，那麼就算沒有這些「正確答案」，似乎也沒有關係了，這對於做這方面研究或產業的人也是一大福音。

在讀這篇論文前對於該如何做到這件事情沒有什麼頭緒，只覺得這是一件非常不簡單的事情，花了不少時間讀完並理解後果然跟預想的一樣複雜，初始化與後續三個階段的優化牽扯到了很多dsp課堂上曾提到過的演算法與資料結構，非常不好懂，而很多英文又難以找到對應的中文意思，又或者英文的構句與中文差異甚遠，所以在解讀上也遇到了不少困難。不過仔細的讀完後也吸收了很多東西，初始化的重要性，為什麼三個階段是這樣設計，漸進式調整參數的好處等等，真心覺得教授跟學長非常厲害，能夠想出這樣的方法並設計實驗去驗證，讓我領教到數位語音處理的博大精深。